ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 8, Issue 4, July 2019

A Novel Sentiment Polarity and Prediction Enhanced Machine Learning Method to Improve Customer Relationship

Amit Kumar, Neetesh Kumar Gupta

Abstract*— The micro-blogging and social network sites are considered as one of the best source of information for the reason that people discuss and freely share their sentiments about a certain subject freely. Posts are generally composed of poorly structured, incomplete, and noisy sentences, irregular expressions, non-dictionary terms, and ill-formed words. The ratings as well as comments related to products are mismatched in most of the posts. Machine learning is about forecasting the future based on the past information. Opinion mining is investigation of reviewer's attitudes, emotions as well as emotions concerning related to events, individuals, topics or issues. SVM is a supervised machine learning procedure which can be applied for classification. The proposed method provides automatically preprocessing of data, words extraction from customers reviews by machine learning. The proposed algorithm used improved Support Vector Machine method for better opinion mining and accurate prediction. The proposed method resolved semantic analysis and opinion mining problems to solve different sentiment polarity. The experimental outcome demonstrate that system is well suitable for opinion polarity and prediction.

Keywords: **Sentiment analysis, Support Vector Machine, Opinion mining, Sentiment Polarity, Machine learning, Social Media.**

I. INTRODUCTION

Twitter has currently become a platform for individuals and organizations who have a strong political, social, or economic concern in enhancing and maintaining their reputation. Twitter, with more than 350 million monthly active users and over 520 million tweets per day. Opinion mining (OM) [1] provides these organizations the capability for monitoring dissimilar posts from social media. Opinion mining is the process of automatically identifying whether a post segment contains opinionated or emotional content, and it can likewise determine the post's polarity. Opinion mining classification aims to categorize the opinion polarity of a tweet as negative, positive, or neutral. Posts are generally composed of poorly structured, incomplete, and noisy sentences, irregular expressions, non-dictionary terms, and ill-formed words. Preprocessing means removing URLs, removing stop words, and replacing negations from users post. A series of pre-processing are applied to reduce the amount of noise in the posts before feature selection. Pre-processing is accomplished comprehensively in existing methodologies, specifically in machine learning-based methods.

Opinion Mining also termed as Sentiment Analysis (SA) is the analysis of public's attitudes, product related opinions, and sentiments or polarity concerning products. The object signifies events, individuals, as well as topics related to products. These subjects may concealed by analyses. The two expressions OM or SA are express a common meaning and can be interchangeable. Opinion Mining is an unending field of research in document mining field. Opinion mining will review different post of users and mine their opinion about related subjects. The Clustering and natural language processing procedure will be applied for opinion mining. A subpart of opinion mining represents by applying natural language processing (NLP) [2] by suggested dissimilar techniques of dictionary for sentimentality analysis of document data as lexicon, specific language dictionary, and corpus. The remaining of the paper is summarized as follows. Section 2 represents related methods. Section 3 provides proposed technique and algorithm. Section 4 provides the implementation details of the proposed work. Section 5 provides conclusions.

II. RELATED WORK

Kampset. el. [3] applied the Word Net dataset to search out the polarity of tokenized words. As they related a target text word to two key words (`bad' and `good') to search the lowest route distance related the pivot words and searched word the in the Word Net dataset hierarchy. The lowest path distance was transformed to an added total and this assessment was stored with the tokenizer word in the words dataset. The described accuracy level of this approach was 63.2%.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 8, Issue 4, July 2019

Littman et.el. [12] mapped the semantic relationship between the search final word and to every dataset word from the designated set of negative and positive words to a real number. By means of subtracting a word's relationship strength to a set of negative words as of its association strength to a group of positive words, an accuracy rate of 83% was accomplished.

Q. Su et. el. Considers the probable of semantic expertise to address these enquiries. Afterward summarizing approaches to disambiguate and extract context information, the author present visualization procedures to discover the geospatial, lexical, and relational background of entities and topics referenced in these sources. The specimens stem as of the, The Climate Resilience Toolkit Media Watch on Climate Change, and the NOAA Media Watch—3 applications that combined environmental resources as of a widespread range of online sources. These schemes not only represents the value of as long as wide-ranging of information the public, but then again also have assisted to improve an innovative communication success metric that goes out there bipolar calculations of sentiment.

SVMs were applied by Li and A. Li [15] as a sentiment polarity classifier. Dissimilar the binary classification difficulty, authors debated that expresser credibility and opinion subjectivity should also be considered into consideration. Authors suggested a framework that make available a condensed numeric summarization of sentiments on micro-blogs platforms. Authors extracted and identified the subjects mentioned in the opinions connected with the requests of users, and then categorized the opinions by using SVM. Authors worked on Twitter posts data for experiment. Author found out that the concern of user opinion subjectivity and credibility is necessary for accumulating micro-blog opinions. The proposed method proved that mechanism can effectually determine market intelligence (MI) for assistant decision-makers by instituting a monitoring method to track exterior opinions on dissimilar aspects of a business in actual time.

Wararat Songpan et.el.[1] Suggests the prediction rating and analysis from customer from hotel examinations who mentioned as open opinion with the help of probability's model as classifier. The suggested classifier models are applied in case study of consumer hotel review's in open posts comments for training dataset to categorized consumer comments as negative or positive. In further step, this classifier model has computed probability which represents value of style to provide the rating by applying naive Bayes procedures, which gives appropriately classifier to 93.47% as comparison with decision tree Methods.

The customer review's from hotel website agent service module for reservation system. The dataset from hotel checkout and checked in logs are arranged for classification. The cleansing of data is performed by removing unused stop data and high frequency word selection by applying classifier model. The consumer reviews which can be negative as well as positive using data from training set and test set.

III. PROPOSED METHOD

The proposed method is divided in three modules. Module one provides data collection from different social media websites like Twitter. Module two provides preprocessing of the different posts gathered from Twitter. Module three provides clustering of all the posts after preprocessing. Opinion mining steps

1. Social media data collection
2. Opinion reviews
3. Opinion identification
4. Opinionative words and phrases
5. Feature selection and extraction
6. Opinion classification
7. Opinion polarity

1. Social media data collection: The first step in opinion mining is to collect large amount of data from social media like Twitter.
2. Opinion review: The second step is to review all the opinion from collected data.
3. Opinion Identification: The next step is to identify the opinion of the users.
4. Opinionative words and phrases: In this step we opinionative all the words and phrases.
5. Feature selection: The next step is feature selection. In opinion classification opinion analysis task is considered an opinion classification problem. The first step in the opinion classification problem is to select and extract text features. Some of the existing features are terms frequency and presence. These features are distinct

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 8, Issue 4, July 2019

word n-grams or words and their frequency sum total. It either gives the words binary weighting. In binary weighting zero if the word appears, or one if otherwise. It also uses term frequency weights to indicate the relative importance of features.

Parts of speech: In parts of speech process finding adjectives, as they are significant indicators of opinions polarity.

Opinion phrases and word: These are words generally used to express opinions comprising bad, or good or hate or like.

Negations: The presence of negative words could change the opinion meaning like not good is comparable to bad.

6. Features: The next step is to extract the required features from the available features.
7. Opinion classification: The next step is to classify the opinion according to requirement.
8. Opinion polarity: The last step is to polarize the opinion.

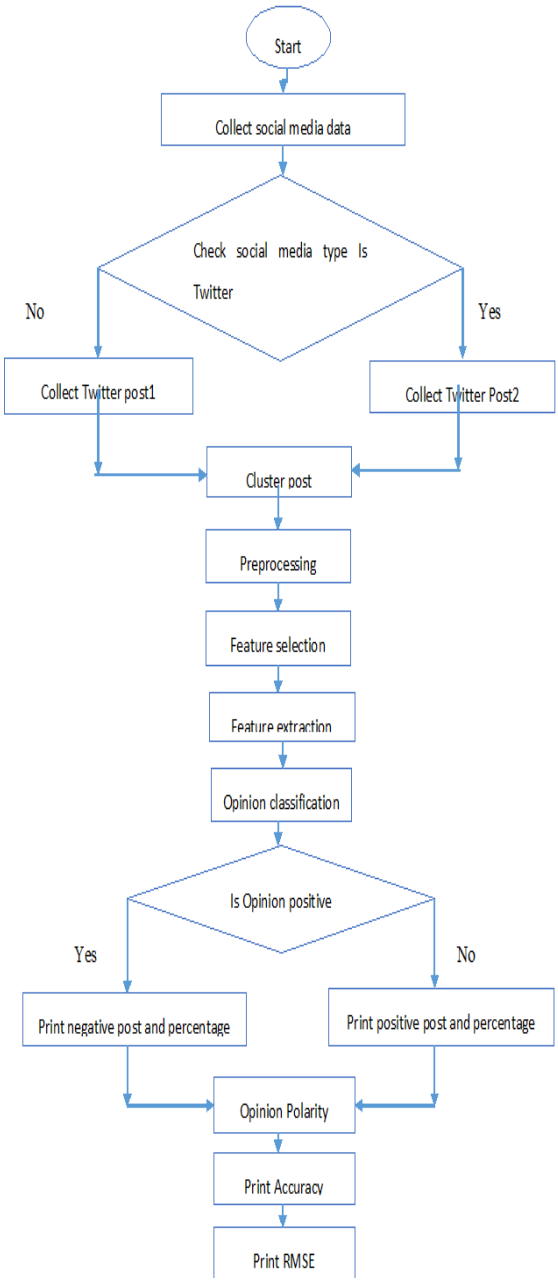

Fig 1: Flow diagram of proposed work

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 8, Issue 4, July 2019

Algorithm: Improved opinion clustering and polarity search algorithm
Input: AP: All posts in dataset, F: Prominent keyword
P: posts, TP: Twitter posts
Output: Positive opinion, negative opinion, sentiment polarity, RMSE, percentage accuracy
Step 1: Collect post from social media like Facebook, Twitter
Step 2: Cluster the data according to P
Step 3: if P==TP then
Apply Twitter API to the post
Apply Google SDK and Opinion mining API
End if
Step 4: Apply preprocessing method
Clearing the text
Removing URL's
Removing Tags
Removing irrelevant contents
Step 5: Apply feature selection and extraction
Prominent keyword extraction
AP= All posts in dataset
Initially F=NULL
For every x in AP do
Y= Extract keyword from posts
For ap_key in posts_keywords do
If ap_key in p_key then
F[ap_key]= F[ap_key] + 1
Else
Insert in K to pa_key
End if
End for
End for
For key in AP do
If AP [key] < threshold then
Remove key from AP
End if
End for
Return AP
Step 6: Apply Support vector machine algorithm for opinion classification
Step 7: Classify positive and negative opinion according to customer P
Step 8: Calculate the opinion polarity
If key_match> 80 then
Print positive accuracy
Else
Print negative accuracy
End if
Step 9: Calculate RMSE
Step 10: Stop

The posts in social media are differ constructed on the opinion status linked with the contents. The total sentiment and opinion related with a post can be neutral, negative or positive. The assessment of opinion at keywords level is made for sentiment analysis which enables posts classify based of attitude of the subject. Primarily the user posts are preprocessed to take out stop words. The next step is keyword extraction. The keywords means the important words in a user post that refer to the subject of related post. The keyword lists linked with entire posts are combined together. The list of final keyword is expressed by filtering out every found keywords that are not as much of used in posts. The algorithm focusing on opinion and polarity associated with common and most keywords being used in every post.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 8, Issue 4, July 2019

The dissimilarity in the opinion scores associated with the found keywords can be a key contributing feature in clustering the related posts constructed on sentiments. The procedure of extracting main prominent keywords is as termed in Algorithm which provides as outcome, the prominent words. The threshold differs dependent on the dataset.

IV. IMPLEMENTATION

For implementation environment used i3 3.0 GHz machine with 4GB RAM. Social media data used for opinion mining and polarity finding. The social media post is collected data from social web sites like Twitter. The different post of users are reviewed and mine their opinion about related subjects.

Dataset

The dataset has been collected from real-world environment of online social networks. The different posts from customers related to different products are collected. Five different groups have been chosen for performing the testing of the proposed work.

Dataset properties

Table 1: Dataset properties

Group	Members	Post
G1	2365	20321
G2	1388	18000
G3	8000	26000
G4	6002	12000
G5	2000	5987

For accessing dataset from Twitter registration is necessary. After registration new project have to be created from https://app.twitter.com. Application management is used to create test app for twitter. For registration some basic information is provided like application name, organization detail, website detail. For authentication Twitter provides keys and access tokens. After getting access tokens and keys we can access the Twitter dataset.

The feature selection is to be characteristics in model that will be take out words from these consumer reviews as words occurred often to 14, 24 and 40 words. There are positive and negative in Table below, which are well-ordered by descendant frequent.

Table 2: Feature selection from frequent words

No.	Words (Positive)	#Frequent	Words (Negative)	#Frequent
1.	Excellent	400	Old	80
2.	Best	380	Inconvenient	80
3.	Better	350	Costly	70
4.	Very Good	320	Not delicious	65
5.	Good	290	Slow	60
6.	Beautiful	280	Expensive	55
7.	Luxurious	270	Troublesome	50
8.	Convenient	250	Problematic	40
9.	Attractive	150	Improve	30
10.	Nice	140	Not beautiful	25
11.	Delicious	130	Immoral	25
12.	Special	120	Not worth	20
13.	Comfortable	100	Uncomfortable	20
14.	Popular	80	Rare	20
15.	Safe	70	Unpleasant	15
16.	Cheap	70	Risky	15
17.	Not expensive	65	Unsafe	15
18.	Thanks	60	Unfriendly	10
19.	new	60	Not good	10
20.	Enjoy	50	Bad	5

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 8, Issue 4, July 2019

The occurrences of frequent words which are positive are observed for attribute modification of individual word of consumer post. The dataset from test and train are separated into three groups, group one formed five negative and five positive words; group two constructed with ten negative and ten positive words and group three is constructed with all negative and positive words.

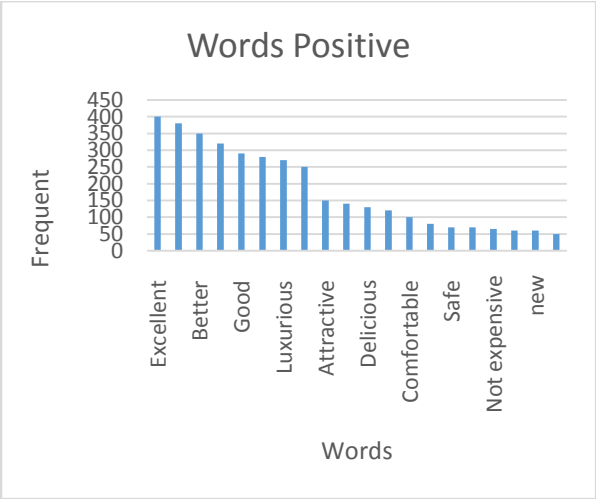

Fig 2: Positive words

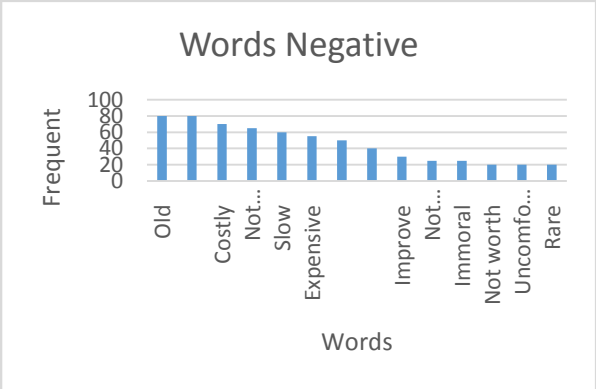

Fig 2: Positive and negative words

Accuracy Analysis

Table 3: Accuracy analysis

	% Accuracy	
Attributes	Naive Base	Proposed
14 words	92.33	93.09
24 words	94.37	95.32
40 words	94.12	95.51
Average	93.61	93.97

The experimental outcomes are tested with open opinions customer reviews of 450 from a twitter and Facebook users posts. The outcomes are compared percentage of precision between naive Bayes and proposed algorithm and dissimilarity the number of feature are take out as 14,24 and 40 words respectively. The accuracy of proposed work is given data that are maximum than naive Bayes all of data groups. Moreover, the maximum of accuracy value is 95.57% with 25 words and also average of proposed algorithm is higher than naive Bayes to 94.37%.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 8, Issue 4, July 2019

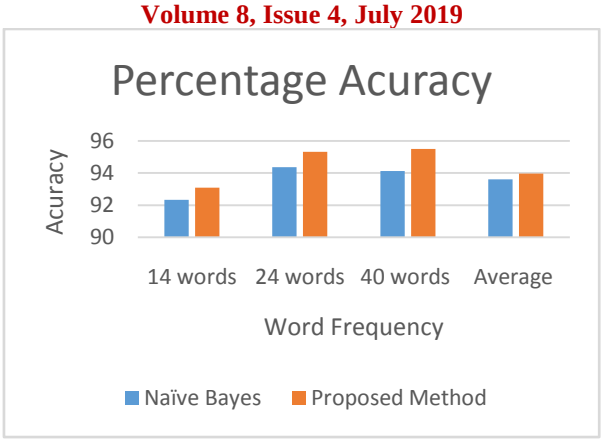

Fig 4: Percentage Accuracy

RMS error

Table 4: RMSE analysis

	RMSE	
Attributes	**Naive Base**	**Proposed**
14 words	**0.3660**	**0.3532**
24 words	**0.2390**	**0.2295**
40 words	**0.2326**	**0.2113**
Average	**0.2792**	**0.2324**

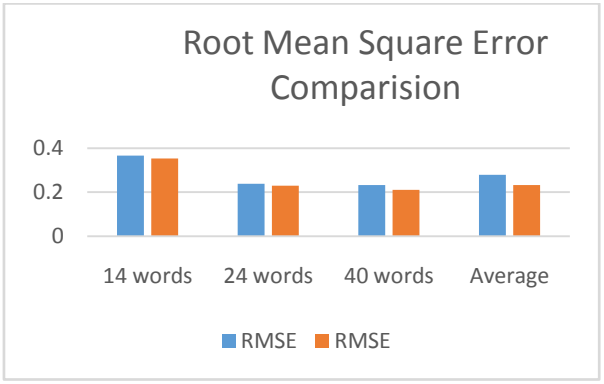

Fig 5: RMSE evaluation

The RMSE values of naïve bayes and proposed method is represented. The table above represents RMSE of data groups. The lowest of RMSE is 40 words testing data that provide rating that are like to actual score from customer review to 0.2113. The rating of 40 words and 14 words are slightly higher value than 40 words to 0.2295 and 0.3532 respectively. The average of proposed method generates rating value that is similar actual rating as 0.2324 and median as 0.2295.

V. CONCLUSIONS

The natural language processing implementations can be applied to facilitate the OM process. With the growing inspiration of online sentiment analysis and reviews on customers, the competence to detect dishonest online appraisals is crucial. It provides improved natural language understanding then can help produce further improved accurate outcomes of OM. In numerous applications, it is significant to consider the context of the text data and the user preferences. The machine learning classification technique Support Vector Machine is used for sentiment polarity. The proposed method improved précised prediction for better opinion mining results. The method provides automatically preprocessing of data and extract related opinion from a sentence. The data is collected from social web sites like Twitter. The different post from users revived and mine their opinion about related subjects. The accuracy is improved to 4% as compared to Naïve Base method. The RMSE is also

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 8, Issue 4, July 2019

reduced to 2 % as compared to Naïve Base method. The proposed novel method can be used for semantic analysis and opinion mining to solve different sentiment polarity. The experimental outcome demonstrates that system is well suitable for accurate sentiment prediction and opinion polarity.

REFERENCES

[1] Waraaraat Songpaan, The Analysis and Prediction of Customer Review Rating Using Opinion Mining, IEEE SEARA 2017, pp. 72-78.

[2] AranoScharil, David Hering, Wallter Rafeels berger, Alexender Hubman-Haidvogeel, Ruslaan Kaamolov, Daniel Fishchl, Michal Foils, and Alberto Weichsebraun, "Semantic System and Visual Tools Applications to Support Environmental Communication", IEEE System Journal, VOL. 12, NO. 2, JUNE 2017, pp. 752-762.

[3] Kampas, J., Marxi, M.K., Mokkeen, R. J.Using Word Nett to Measure Semantic Orientation of Adjective. LRREC 2004. Vol. IV, pp. 1215-1218.

[4] Andrevskia, A., Berglar, S., Urseana, M.All Blogs Are Not Made Equals: Exploring Genre Differences in Sentiment Tagging of Blog. International Conference on Weblog and Social Media (ICSWM-2007), Booulder, CAO. 2007.

[5] Vandaana V. Chaudhary, Chitraa A. Dhawala and Sanjeev Mishra, "Sentiment Analysis Classification: A Brief Review", I J C T A, 9(23) 2016, pp. 447-454.

[6] Ahni-DangVao, Quangi- Phauca Nagyen, and Chel-Young OCKI, "Opinion Aspect Relationship in Cognizing Customer Feeling through Reviews", IEEE 2017, pp. 5315-5327.

[7] Athiroo U, and Saboo M. Thampl, "Linguistic Feature Constructed Filtering Mechanisms for Recommendation Post in a Social Networking Group", IEEE 18, pp. 4479-4494.

[8] S. 1. Wu, R.D. Chiang and Z.H. Ji, Development of a Chinese opinion mining system for application to Internet online forum, The Journal of Supercomputing, Springer US[Online], 2016.

[9] L.Liu, Z. Li, and Ci.Li, Analysis of customer satisfaction from China reviews by using opinion mining, proceeding of the 7th IEEE International Conferences on Software Engineering and Service Science (ICESES). 2015, pp.105-109.

[10] X.Xu, Z.Guo,Q.Su, H.Guo, X. Wu, X. Zhang and Bi .Sweni. Hidden Sentiment association in Chinese web opinion mining. Proceeding of the 17th International Conference on WWW, 2008, pp.969-978.

[11] R.M. Duwaii and I. Qaarqaz, Arabic Sentiment Analysis using Supervised Classification. Proceeding of 2014 International Conference on Internet of Things and Cloud Computing. 2014, pp. 589-593.

[12] T.V. Le, H.S. Le, and V.T. Phaami, Aspect Analysis for Opinion Mining of Vietnamese Text. Proceeding of International Conference on Advance Application and Computing, 2015, pp.128-133.

[13] D.C. Londhee and B.V. Raaut, "Survey on opinion mining and summarization of user review on web", Journal of Information Technology and Computer Science, Volume. 4, 2014, pp. 1036-1040.

[14] S. Mohammed, Fiaidhi, O. Mohammed, T.H, Kim, S. Fong, Opinion Mining over Twitter and space: Classifying tweets programically using the R approach. Proceeding of the 7th International Conference on Digital Info Management, 2012, pp. 323-329.

[15] R. Zhang, Li, W, L. Lin, Yu and C. Sun, Opinion mining and sentiment analysis in social nets: A retweeting structure with aware approach. Proceeding of the 6th International Conferences on Cloud Computing and Utility, 2014, pp.880-885.